A. A. Rogov[1], *A. G. Varfolomeyev*[1], *A. O. Timonin*[1], *K. A. Proença*[2]

A PROBABILISTIC APPROACH TO COMPARING THE DISTANCES BETWEEN PARTITIONS OF A SET*

[1] Petrozavodsk State University, 33, Lenin pr., Petrozavodsk, 185910, Russian Federation
[2] Feedzai, Avenida D. Joao II, Lote 1.16.01 Piso 11, Lisbon, 1990-083, Portugal

This article describes and compares a number of classical metrics to compare different approaches to partition a given set, such as the Rand index, the Larsen and Aone coefficient, among others. We developed a probabilistic framework to compare these metrics and unified representation of distances that uses a common set of parameters. This is done by taking all possible values of similarity measurements between different possible partitions and graduating them by using quantiles of a distribution function. Let λ_α be a quantile with α level for distribution function $F_\rho(t) = P(\rho < t)$. Then if the proximity measurement ρ is not less than λ_α, we can conclude that $\alpha \cdot 100\%$ of randomly chosen pairs of partitions have a proximity measurement less than ρ. This means that these partitions can neither be considered close nor similar. This paper identifies the general case of distribution functions that describe similarity measurements, with a special focus on uniform distributions. The comparison results are presented in tables for quantiles of probability distributions, using computer simulations over our selected set of similarity metrics. Refs 9. Table 1.

Keywords: distance between partitions of a set, probabilistic approach, comparing the distances.

А. А. Рогов[1], *А. Г. Варфоломеев*[1], *А. О. Тимонин*[1], *К. А. Проенца*[2]

ВЕРОЯТНОСТНЫЙ ПОДХОД К СРАВНЕНИЮ МЕР БЛИЗОСТИ МЕЖДУ РАЗБИЕНИЯМИ МНОЖЕСТВА

[1] Петрозаводский государственный университет, Российская Федерация, 185910, Петрозаводск, проспект Ленина, 33
[2] Фидзаи (Feedzai), Португалия, 1990-083, Лиссабон, проспект Д. Жуана II, 1.16.01, 11

В статье рассматривается ряд классических метрик (индекс сходства разбиений, предложенный Рандом; коэффициент Ларсена—Аоне и др.) между разбиениями одного множе-

Rogov Aleksandr Aleksandrovisch — doctor of technical sciences, professor, head of department; rogov@psu.karelia.ru
Varfolomeyev Aleksey Gennadievich — PhD of physical and mathematical sciences, associate professor; avarf@petrsu.ru
Timonin Artem Olegovich — postgraduate student; timonin.artem@gmail.com
Proença Kseniya Aleksandrovna — data scientist; kseniya.proenca@gmail.com

Рогов Александр Александрович — доктор технических наук, профессор, заведующий каферой; rogov@psu.karelia.ru
Варфоломеев Алексей Геннадьевич — кандидат физико-математических наук, доцент; avarf@petrsu.ru
Тимонин Артем Олегович — аспирант; timonin.artem@gmail.com
Проенца Ксения Александровна — специалист по обработке и анализу данных; kseniya.proenca@gmail.com

* The work was supported by the Program of Strategic Development of Retrozavodsk State University within the framework of the implementation of a set of activities for the development of research activities for 2012–2016.

Работа выполнена при поддержке Программы стратегического развития Петрозаводского государственного университета в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

 https://doi.org/10.21638/11701/spbu10.2018.102

ства. Унифицированы формулы для их вычисления на основании одинаковых параметров. Разработан вероятностный подход к сравнению приведенных мер близости (сходства). Для этого требуется градуировка интервала возможных значений мер близости между возможными разбиениями с помощью квантилей функции распределения. Пусть λ_α — квантиль уровня α для функции распределения $F_\rho(t) = P(\rho < t)$. Тогда, если мера близости ρ оказывается не меньше, чем λ_α, можно сделать вывод, что $\alpha \cdot 100\%$ случайно выбранных пар разбиений имеют между собой меру близости меньше, чем ρ. Следовательно, их нельзя считать близкими или похожими. Получен общий вид функции распределения для приведенных мер близости. Подробно изучен случай равномерного распределения элемента разбиения в любой группе. Для ряда мер близости приведены таблицы квантилей функции распределения, которые были построены с помощью компьютерного моделирования. Библиогр. 9 назв. Табл. 1.

Ключевые слова: меры близости между разбиениями множеств, вероятностный подход, сравнение мер близости.

Introduction. The numerical comparison of disjoint set partitions, which we call clusters, is a well studied subject in the literature [1–5]. We consider three types of similarity measurements between clusters, following Meilă [1]:

1) by checking if a given object belongs or not to each known cluster [2, 3];
2) by comparing clusters regarded as sets [4, 5];
3) by calculating the delta produced from moving an object between two partitions [1].

There is, however, no approach to precisely compare these similarity measurements, since each of them has its own advantages and disadvantages [6]. This paper takes a step forward in this direction, comparing and relating different similarity measurements.

Similarity measurements of set partitions. Assuming we have a set with n elements and two disjoin non-empty partitions (clusters) of this set. Let's call m_{jl} a frequency of elements to belong to clusters with numbers j and l in the first and second partitions. The paper [1] proposes to express all measurements that compare such kind of subsets through these frequencies m_{jl}. Let's call m_{j*} and m_{*l} as marginal frequencies. Their values will be equal to the number of elements in the clusters with numbers j and l mentioned above with numbers j and l. The following relations stay for the frequencies introduced above:

$$\sum_{j,l} m_{jl} = n, \quad \sum_{l} m_{jl} = m_{j*}, \quad \sum_{j} m_{jl} = m_{*l}.$$

Calling matrix M a matrix consisting of the elements m_{jl}. Then for two identical partitions in each row j and in each column l of the matrix M will be only one non-zero element on the main diagonal. When using additional values

$$T = \sum_{j,l} m_{jl}^2 - n, \quad S = \sum_{j} m_{j*}^2 - n, \quad Q = \sum_{l} m_{*l}^2 - n$$

then, in these terms, a partition similarity index, proposed by Rand [2], will be equal to

$$R = 1 - \frac{S + Q - 2T}{n(n-1)} \tag{1}$$

and the proximity factor introduced in [3] appears as

$$F = \frac{T}{\sqrt{SQ}}.$$

Let μ_i be some proximity measurement between two clusters from different partitions that contain the i-th element of the set. We call ρ_k the mean of proximity measurements

between two partitions of a set splitter into k non-empty subsets, which is calculated by formula

$$\rho_k = \frac{\sum_{i=1}^{n} \mu_i}{n}. \tag{2}$$

The coefficient ρ_k is calculated in the following way: each pair of clusters j and l from the first and second partitions is being compared as many times as the pair has common elements.

Introducing the notation $\mu(j,l)$ for the proximity coefficient between clusters j and l allow us to write the proximity coefficient (2) as

$$\rho_k = \frac{1}{n} \sum_{j,l} m_{jl} \cdot \mu(j,l).$$

Any of the proximity coefficient can be used as a measurement $\mu(j,l)$ between two sets [7], or, what is the same, two (0,1) vectors [8]. For example:

$$\mu^1(j,l) = \frac{m_{jl}}{m_{j*} + m_{*l} - m_{jl}},$$

$$\mu^2(j,l) = \frac{m_{jl}}{\max(m_{j*} + m_{*l})}, \tag{3}$$

$$\mu^3(j,l) = \frac{2m_{jl}}{m_{j*} + m_{*l}}.$$

Obviously, $\mu(j,l)$ takes values from 0 to 1. The more matching elements are in sets, the closer these coefficients are to 1. Look like:

$$\rho_k^1 = \frac{1}{n} \sum_{j,l} \frac{m_{jl}^2}{m_{j*} + m_{*l} - m_{jl}}\ , \tag{4}$$

$$\rho_k^2 = \frac{1}{n} \sum_{j,l} \frac{m_{jl}^2}{\max(m_{j*} + m_{*l})}, \tag{5}$$

$$\rho_k^2 = \frac{1}{n} \sum_{j,l} \frac{2m_{jl}^2}{m_{j*} + m_{*l}}. \tag{6}$$

The proposed similarity coefficient is the proximity measurement of weighted sum between all clusters from the first and second partitions. The corresponding intersection cardinalities are used as weights. The coefficients from the papers [4, 5] are the most similar to the proposed measurement. They are also calculated using a pairwise comparison of clusters, but the summation is unweighted and is not performed for all pairs of clusters. For example, the Larsen—Aone coefficient [5] is

$$L = \sum_j \max_l \mu^3(j,l).$$

Proximity measurements comparison. Regardless of which proximity measurements are used, the problem arises when determining which measurements values can be considered large (close to 1) and which ones should be considered small. The solution to this problem will answer the question if the difference between the partitions is significant

or appeared to be random. This article develops an approach relaying on a probabilistic model for generating partitions and is based in the previously described work from [7, 9]. This approach allows us to set a specific value — threshold to determine "big" and "small" values of the similarity measurement. If the value of the proximity measurements seldom appears to be the same or higher then it is considered "large". The opposite is also true: if values occur frequently, it is considered to be "small". The paper proposes a method for constructing quantitative estimates for the concepts "rarely" and "often" based on the probability distribution of the proximity measurements values.

We perform a random experiment that generates a pair of partitions. We also introduce a probability measurement for the set of outcomes of the experiment. Like this we obtain a probability distribution of the proximity measurements values. This lets us to perform a calibration of the possible values range using quantiles of the proximity measurements distribution function. Let λ_α be a quantile with α level for distribution function $F_\rho(t) = P(\rho < t)$. Then if the proximity measurement ρ is not less than λ_α, we can conclude that $\alpha \cdot 100\%$ of randomly chosen pairs of partitions have a proximity measurement less than ρ. A similar approach was considered in the paper [9] when comparing distances between subsets, and in the paper [7] when comparing dendrograms.

The distance probability distribution in a general case. Let $U = \{u_1, u_2, \ldots, u_n\}$ be a set of n elements. X and Y are the two of its partitions, both consisting of k groups. We represent the partitions X and Y in the form of vectors x and y of dimension n. X is constructed according to the principle: $x_i = j$ if and only if u_i belongs to the jth group. Y is being built in the same way. We name $p_j^i,\ i = 1, \ldots, n,\ j = 1, \ldots, k$, the probability of appearance of the element u_i in the j-th group. Then we can consider a random experiment that consists of n independent tests and in each test the element u_i can appear in any partition group. It appears that each test can have k^2 species $A^{ijl} = \{x_i = j, y_i = l\}$, where i is the test number. Let $I(A)$ be the indicator of the event A. Then $m_{jl} = \sum_{i=1}^{n} I\left(A^{ijl}\right)$, $m_{j*} = \sum_{l=1}^{k}\sum_{i=1}^{n} I\left(A^{ijl}\right)$, $m_{*l} = \sum_{j=1}^{k}\sum_{i=1}^{n} I\left(A^{ijl}\right)$.

We construct the set of events $B = \left(A^{111}, A^{112}, \ldots, A^{nkk}\right)$, taking into account the condition that empty groups are not allowed. Each element can be exactly in one group in each of the two partitions. Then for each $i \in 1, \ldots, n$ the condition $\sum_{j=1}^{k}\sum_{l=1}^{k} I\left(A^{ijl}\right) = 1$ is right, and the condition $\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=1}^{k} I\left(A^{ijl}\right) = n$ is true.

If $I\left(A^{irt}\right) = 1$ is true, then for any $j \neq r,\ l \neq t$ appears $I\left(A^{ijl}\right) = 0$. There are $(k-1)^2$ of such pairs of j and l, therefore $\forall i \in 1, \ldots, n : \sum_{j=1}^{k}\sum_{l=1}^{k}\left(1 - I\left(A^{ijl}\right)\right) = (k-1)^2$.

To guarantee the absence of empty groups, we introduce the conditions $\forall j : \sum_{i=1}^{n}\sum_{l=1}^{k} I\left(A^{ijl}\right) \geqslant 1,\ \forall l : \sum_{i=1}^{n}\sum_{j=1}^{k} I\left(A^{ijl}\right) \geqslant 1$. Thus, the set B of outcomes follows the conditions:

$$B = \left\{ \begin{array}{c} \left(A^{111}, A^{112}, \ldots, A^{nkk}\right) \\ \sum_{j=1}^{k}\sum_{l=1}^{k} I\left(A^{ijl}\right) = 1 \\ \sum_{j=1}^{k}\sum_{l=1}^{k}\left(1 - I\left(A^{ijl}\right)\right) = (k-1)^2 \\ \forall j : \sum_{i=1}^{n}\sum_{l=1}^{k} I\left(A^{ijl}\right) \geqslant 1, \quad \forall l : \sum_{i=1}^{n}\sum_{j=1}^{k} I\left(A^{ijl}\right) \geqslant 1 \end{array} \right\}.$$

As it was shown in the previous sections, different similarity coefficients between two set partitions are described as functions of m_{jl}, i. e. $p(X, Y) = h(m_{jl})$. To calculate the distribution function of the random variable $p(X, Y)$, we can use the formula for conditional probabilities. We call H the event that the partition does not contain empty groups. Then

$$P\left(p\left(X,Y\right)<t\mid H\right)=\frac{P\left(\left\{p\left(X,Y\right)<t\right\}\cdot H\right)}{P\left(H\right)},$$

$$P\left(\left\{p\left(X,Y\right)<t\right\}\cdot H\right)=\sum_{m_{jl}\in C}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\prod_{j=1}^{k}\prod_{l=1}^{k}\prod_{i=1}^{n}\left(p_j^i p_l^i\right)^{I\left(A^{ijl}\right)},$$

where

$$C=\left\{m_{jl}\in Z:m_{jl}\geqslant 0,\ \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=1}^{k}m_{jl}=n,h\left(m_{jl}\right)<t\right\};$$

$$P\left(H\right)=\ P\left(\left\{p\left(X,Y\right)<\infty\right\}\cdot H\right).$$

Then the distribution function of the random variable $p\left(X,Y\right)$ can be written in the following form:

$$F\left(t\right)=P\left(\left\{p\left(X,Y\right)<t\right\}\cdot H\right)=\frac{1}{P\left(H\right)}\sum_{m_{jl}\in C}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\prod_{j=1}^{k}\prod_{l=1}^{k}\prod_{i=1}^{n}\left(p_j^i p_l^i\right)^{I\left(A^{ijl}\right)}. \tag{7}$$

Let us consider a special case.

A special case. Uniform distribution. Let's call $p_j^i=\frac{1}{k}$. Then the formula (7) takes the form

$$F_p\left(t\right)=\frac{1}{P\left(H\right)}\sum_{m_{jl}\in C}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\prod_{j=1}^{k}\prod_{l=1}^{k}\prod_{i=1}^{n}\left(\frac{1}{k}\right)^{2I\left(A^{ijl}\right)}=$$

$$=\frac{1}{P\left(H\right)}\sum_{m_{jl}\in C}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\prod_{i=1}^{n}\left(\frac{1}{k}\right)^{2}=\frac{1}{P\left(H\right)}\sum_{m_{jl}\in C}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\left(\frac{1}{k}\right)^{2n},$$

$$P\left(H\right)=\sum_{m_{jl}\in C_1}\sum_{\left(A^{111},\ldots,A^{nkk}\right)\in B}\left(\frac{1}{k}\right)^{2n},$$

where

$$C_1=\left\{m_{jl}\in Z:m_{jl}\geqslant 0,\ \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=1}^{k}m_{jl}=n\right\}.$$

We developed the program to calculate the quantiles of the proximity measurement distributions using the formulas (3)–(5). The calculation we done using simulation modelling. The table presents the calculated quantile values of some values n and k with different α. In a rectangle corresponding to the same values of n and k, the measurements were calculated with formulas (1), (4)–(6). The upper left corner values were calculated using the formula (4); the upper right corner — the formula (5); the lower-left corner — formula (6) and the lower right corner was computed using the formula (1). The number of experiments was 10 000.

Table. **Proximity measurements quantities for different α**

α	$k\backslash n$	5		10		30	
0.2	3	0.307	0.333	0.229	0.313	0.208	0.299
		0.467	0.400	0.370	0.422	0.343	0.508
	5	0.0	0.0	0.252	0.298	0.146	0.215
		0.0	0.0	0.397	0.600	0.252	0.623
	7	0.0	0.0	0.442	0.450	0.146	0.205
		0.0	0.0	0.582	0.778	0.250	0.703
0.1	3	0.307	0.333	0.229	0.313	0.209	0.300
		0.467	0.400	0.370	0.422	0.343	0.508
	5	0.0	0.0	0.254	0.292	0.147	0.216
		0.0	0.0	0.402	0.600	0.253	0.621
	7	0.0	0.0	0.442	0.450	0.146	0.206
		0.0	0.0	0.583	0.778	0.252	0.706
0.05	3	0.307	0.333	0.229	0.313	0.209	0.300
		0.467	0.400	0.370	0.422	0.344	0.508
	5	0.0	0.0	0.257	0.292	0.148	0.216
		0.0	0.0	0.399	0.600	0.252	0.623
	7	0.0	0.0	0.433	0.450	0.146	0.206
		0.0	0.0	0.582	0.778	0.252	0.706

Conclusions. The introduced proximity measurements estimation allows us to evaluate values obtained on different sets and using different proximity measurements. For example, we assume that the similarity measurements $\rho^1(X,Y)$ and $\rho^2(A,B)$ are statistically close with precision $\varepsilon > 0$ if the inequality $\left|F_\rho^1\left(\rho^1(X,Y)\right) - F_\rho^2\left(\rho^2(A,B)\right)\right| \leqslant \varepsilon$ is true.

References

1. Meilă M. Comparing clusterings by the variation of information. *Learning Theory and Kernel Machines. Lecture Notes in Computer Science* (Springer), 2003, vol. 2777, pp. 173–187.

2. Rand W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971, vol. 66, pp. 846–850.

3. Fowlkes E. B., Mallows C. L. A Method for comparing two Hierarchical Clusterings. *Journal of the American Statistical Association*, 1983, vol. 78, pp. 553–569.

4. Meilă M., Heckerman D. An experimental comparison of model-based clustering methods. *Machine Learning*, 2001, vol. 42, pp. 9–29.

5. Larsen B., Aone C. Fast and effective text mining using linear time Document Clustering. *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 1999, pp. 16–22.

6. Steel M. A., Penny D. Distributions of tree comparison metrics — Some new results. *Systematic Biology*, 1993, vol. 42, pp. 126–141.

7. Sidorov Y. V., Kirikov P. V., Rogov A. A. Sravnenie dendrogramm s ravnyim chislom vershin [Dendrograms comparison with an equal vertices number]. *Scientific notes of Petrozavodsk State University. Series Natural and Technical Sciences.* Petrozavodsk, Petrozavodsk State University Publ., 2011, no. 8, pp. 108–110. (In Russian)

8. Warrens M. J. On Robinsonian dissimilarities, the consecutive ones property and latent variable models. *Advances in Data Analysis and Classification*, 2009, vol. 3, pp. 169–184.

9. Varfolomeyev A. A., Kirikov P. V., Rogov A. A. Veroyatnostnyiy podhod k sravneniyu rasstoyaniy mezhdu podmnozhestvami konechnogo mnozhestva [Probabilistic approach to distances comparison between subsets of a finite set]. *Scientific notes of Petrozavodsk State University.* Petrozavodsk, Petrozavodsk State University Publ., 2010, no. 8, pp. 83–88. (In Russian)

For citation: Rogov A. A., Varfolomeyev A. G., Timonin A. O., Proença K. A. A probabilistic approach to comparing the distances between partitions of a set. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, iss. 1, pp. 14–19. https://doi.org/10.21638/11701/ spbu10.2018.102

Статья рекомендована к печати проф. А. П. Жабко.

Статья поступила в редакцию 7 октября 2017 г.

Статья принята к печати 11 января 2018 г.